D654–D659 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 7 November 2013*
doi:10.1093/nar/gkt1048

DOOR 2.0: presenting operons and their functions through dynamic and integrated views

Xizeng Mao[1,2], Qin Ma[1,2], Chuan Zhou[1,3], Xin Chen[1,4], Hanyuan Zhang[1,4], Jincai Yang[5], Fenglou Mao[1], Wei Lai[1] and Ying Xu[1,2,4,*]

[1]Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA, [2]BioEnergy Science Center (BESC), Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA, [3]School of Mathematics, Shandong University, Jinan, Shandong 250100, China, [4]College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China and [5]College of Computer Science, Central China Normal University, Wuhan, Hubei 430079, China

Received September 1, 2013; Revised October 10, 2013; Accepted October 11, 2013

ABSTRACT

We have recently developed a new version of the DOOR operon database, DOOR 2.0, which is available online at http://csbl.bmb.uga.edu/DOOR/ and will be updated on a regular basis. DOOR 2.0 contains genome-scale operons for 2072 prokaryotes with complete genomes, three times the number of genomes covered in the previous version published in 2009. DOOR 2.0 has a number of new features, compared with its previous version, including (i) more than 250 000 transcription units, experimentally validated or computationally predicted based on RNA-seq data, providing a dynamic functional view of the underlying operons; (ii) an integrated operon-centric data resource that provides not only operons for each covered genome but also their functional and regulatory information such as their *cis*-regulatory binding sites for transcription initiation and termination, gene expression levels estimated based on RNA-seq data and conservation information across multiple genomes; (iii) a high-performance web service for online operon prediction on user-provided genomic sequences; (iv) an intuitive genome browser to support visualization of user-selected data; and (v) a keyword-based Google-like search engine for finding the needed information intuitively and rapidly in this database.

INTRODUCTION

Operons have been widely used as the basic transcriptional and functional units when studying higher-level functional systems in prokaryotes such as biochemical pathways, networks and regulation systems since the concept was proposed by French scientists Jacob and Monod in 1960 (1). Although it has never been suggested by the two scientists in their original paper, computational prediction of operons often treats them as units that do not overlap with each other (2,3), as this greatly simplifies operon prediction on the genomic scale. For the past decade, an increasingly popular term being used is 'transcriptional units', which are experimentally identified 'operons' as defined by Jacob and Monod in 1960 and may have overlaps.

The emergence of large-scale RNA-seq data for increasingly more prokaryotic organisms has made it possible to elucidate 'operons' in their full complexities, as few genome-scale transcriptomic data collected under multiple conditions have been used to reveal the dynamic structures of the statically predicted operons under different experimental conditions (4). We envision that the need for elucidation of the condition-dependent transcriptional units (TUs) (4,5) will continue to increase, as increasingly more RNA-seq data become available. Throughout this article, we use operons to refer to static non-overlapping 'transcriptional units' while using TUs to refer to operons according to the original definition of Jacob and Monod, i.e. sequences of consecutive genes that each encode a single RNA molecule along with their own promoters and terminators. The typical relationship between operons and TUs is that TUs tend to be sub-units of operons, while in some cases, a TU may span more than one operon.

As of now, a number of operon databases have been publicly deployed by different research groups, including RegulonDB (5), ODB (6), DBTBS (7), OperonDB (8), ProOpDB (9) and DOOR (10) that was developed by our laboratory. These databases differ in their coverage of the operon information, and only a few have TU data. For example, the current version of RegulonDB contains

*To whom correspondence should be addressed. Tel: +1 706 542 9779; Fax: +1 706 542 9751; Email: xyn@bmb.uga.edu

The authors wish it to be known that, in their opinion, the first four authors should be regarded as Joint First Authors.

>800 unique TUs for *Escherichia coli* (5) and ODB has 10 000 TUs (11), both collected from the public domain. Most of these databases do not contain regulatory information for their operons such as transcription factor binding sites and transcription terminators. In addition, none of these database servers provide services for online operon prediction on user-provided genomic sequences; only ODB provides 4812 reference operons that can potentially be used to assist operon prediction.

The new version of the DOOR database, DOOR 2.0, covers all the 2072 completely sequenced prokaryotic genomes in the NCBI genome database (as of April 2012), which is three times the number of genomes covered in its previous version published in 2009. In addition, DOOR 2.0 has several new features, namely, (i) 254 685 TUs collected from public databases such as RegulonDB (5) and Palsson's dataset (4) or computationally predicted based on RNA-seq data; (ii) an integrated operon-centric data resource offering operons, their regulatory binding sites for transcription initiation (TFBSs), transcription terminators, gene-expression levels estimated based on RNA-seq data and their conservation information across multiple genomes; (iii) a high-performance web service for operon prediction on user-provided genomic sequences, powered by a backend computer cluster with >150 computing nodes; (iv) an intuitive genome browser to support visualization of user-specified data in the database; and (v) a keyword-based Google-like search engine for finding the needed information in the database intuitively and rapidly. To the best of our knowledge, DOOR 2.0 is the first web-based operon database that integrates all such capabilities. Together, it provides an easy-to-use environment for discovering new information and synthesizing new knowledge about operons, their function, regulation and evolution across all sequenced prokaryotes. The database can be accessed at http://csbl.bmb.uga.edu/DOOR/, which will be updated on a regular basis when new prokaryotic genomes are released.

DATABASE UPDATE

DOOR 2.0 contains operons for 2072 complete prokaryotic genomes that were downloaded from the NCBI Genome FTP server (April 2012), which consists of 1939 bacteria and 133 archaea, with 2205 chromosomes and 1645 plasmids. We predicted 1 323 902 multi-gene operons using our prediction program (12), on average ~583 such operons per chromosome and ~24 operons per plasmid, and 2 578 949 single-gene operons, as detailed in Table 1. All the operons are stored in a MySQL relational database on our server and can be accessed efficiently through different ways. A user can browse operons by organisms or chromosomes/plasmids that are organized into a searchable HTML table under the 'Organisms' navigation menu. The operons for an organism can be downloaded through the 'Download' link on the 'listing operons' page. A user can search for individual operons in the search box using keyword(s), which is located in the upper right corner of the web page (Figure 1A). The user can also specify more complex queries by using multiple keywords connected through Boolean operators just as in Google, whose details can be found in the online manual at the DOOR 2.0 web server (Figure 1A).

NEW FEATURES IN DOOR 2.0

DOOR 2.0 consists of 254 685 (1385 experimentally verified and 253 300 predicted) TUs for 24 prokaryotic genomes, 6408 verified TFBS for 203 prokaryotic genomes, 3 456 718 *Rho*-independent terminators for 2072 genomes and 6 975 454 conserve operons. The reason that only 24 organisms have TU information is that only those organisms each have a large number of RNA-seq data, sufficient for reliable TU predictions. We expect that this number will increase rapidly as the more genome-scale RNA-seq data become available.

The previous version of DOOR supports the following features: (i) an online operon database for 675 prokaryotic genomes, (ii) a menu-based interface for finding user-specified attributes of operons, (iii) a motif prediction service for user-specified operons and (iv) a Wiki page to facilitate communications between the users and the developer. DOOR 2.0 has kept all these features except for 'operon search based on its number of genes' and the Wiki page, as we found that they have not been actively used based on the usage statistics in the past 4 years. In addition, DOOR 2.0 has a number of new features, selected based on users' inputs as well as our expectation of what might be needed by users of an operon database, based on our own research experience of comparative genome analyses.

INTEGRATION OF TUs

An operon may be transcribed into different TUs under different experimental conditions, which tend to be sub-operonic with their own promoters and/or terminators (13),

Table 1. The key statistics of DOOR 2.0

Category	Number of operons	With TUs	With TFBSs	With terminators	Number of conserved operon
Species	2072	24	203	2072	N/A
Chromosome	2205	24	224	2205	N/A
Plasmid	1645	0	13	1645	N/A
Operon (M)	1 323 902	254 685	4229	1 493 272	6 975 454
Operon (S)	2 578 949	N/A	2260	1 963 446	N/A

Operon (M), multi-gene operons; Operon (S), single-gene operons; N/A, not applicable.

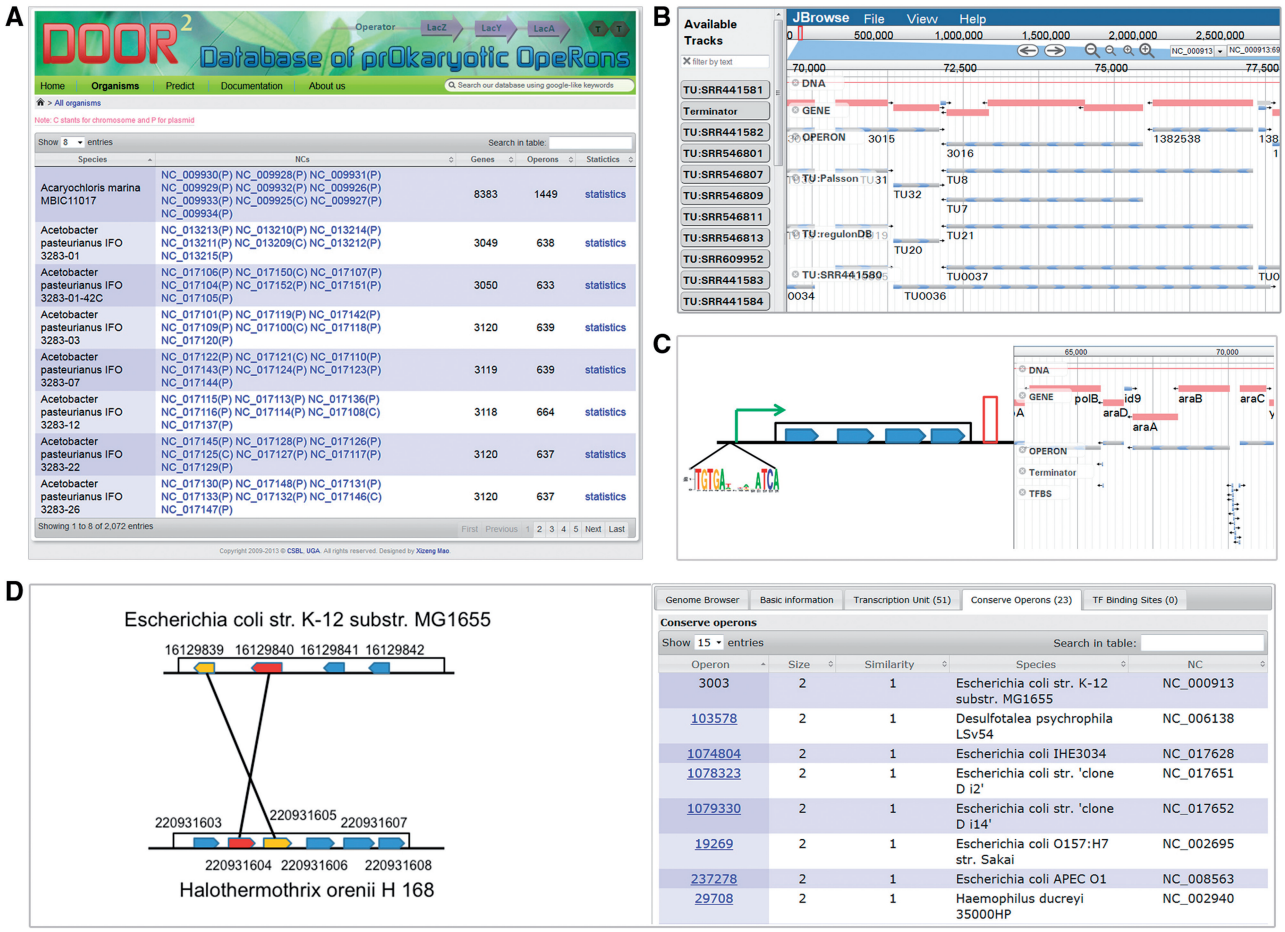

Figure 1. (**A**) A screenshot of a display window. (**B**) A display of TUs, with the red bars representing genes, the first row of the blue bars representing multi-gene operons and the following rows of blue bars being TUs under different conditions. (**C**) A display of validated or predicted transcription factor binding sites (the left bottom) and *Rho*-independent terminators (on the right); and (**D**) conserved operons.

whereas in some cases could be super-operonic, which spans at least two operons (4). The TUs can be derived through RNA-seq analysis. In addition, numerous TUs have been experimentally validated in various prokaryotes and stored in public databases (5).

We have collected 1385 experimentally validated TUs in *E. coli* from the RegulonDB database (5) and Palsson's dataset (4), with 941 and 842 from the first and the second dataset, respectively. In addition, we have predicted 253 300 TUs for all 24 bacterial genomes with genome-scale RNA-seq data in the NCBI SRA database (release of March 2013) (14) using our in-house program *SeqTU* (manuscript in preparation), 119 RNA-seq datasets being used for our prediction. *SeqTU* is a machine learning-based classifier for detecting boundaries between consecutive TUs on the same strand of a genome.

All the TUs are stored in a relational database and can be retrieved and displayed through the genome browser (Figure 1B). A user can examine TUs within an operon using the 'operon' page. Like operons, each TU has its own gene list with their genomic coordinates, underlying RNA-seq data, and an accuracy score if the TU is predicted by *SeqTU*. These items are individually clickable for more detailed information. A user can examine individual TUs via the genome browser by double-clicking the relevant RNA-seq ID in the left panel of the browser, which are not displayed by the default setting. To help the users to examine the expression values of a gene of interest, DOOR 2.0 provides a BigWig XY plot for each underlying RNA-seq data (15), where a user can double-click on the relevant BigWig item for more detailed information.

INTEGRATION OF TRANSCRIPTION REGULATORY ELEMENTS

DOOR 2.0 provides experimentally verified TFBSs for 203 organisms and predicted intrinsic transcriptional terminators for all 2072 organisms, which can be used to study transcriptional regulation of operons.

We have collected 6489 verified TFBS for 203 organisms from *RegulonDB* (for *E. coli* only) (5) and RegTransBase (for 202 organisms) (16). All the TFBSs for each operon, if available, are displayed in an HTML table on the *operon* page, and can be examined along the underlying chromosome through the genome browser. TFBSs are not shown by default when an operon is displayed, but a user can double-click on the relevant menu

in the left panel of the genome browser to turn on this feature. Each TFBS displayed is clickable, through which a user can find out the more detailed information such as its name, genomic coordinates and the DNA sequence (see Figure 1C).

DOOR 2.0 also provides a *de novo* TFBS prediction capability for user-selected operons using two programs: BoBro (17,18) and MEME (19,20). In all, 300-bp upstream sequences of the selected operons will be automatically retrieved from the selected genomes, and the predicted TFBSs will be displayed in an HTML table along with the coordinates, the *P*-value measuring the statistical significance of the prediction, the consensus sequence and a WebLogo (21) (see Figure 2).

It is known that prokaryotes use two different mechanisms of transcription termination: *Rho*-independent (intrinsic) and *Rho*-dependent (22). *Rho*-dependent termination involves the binding of a *Rho* factor to the mRNA to destabilize the RNA–DNA interaction to stop transcription, whereas *Rho*-independent termination functions by creating an RNA hairpin loop to stop the RNA polymerase (23). *Rho*-independent terminators can be reliably predicted based on identification of the conserved RNA hairpin loop, whereas *Rho*-dependent terminators cannot yet, due to the lack of known signals, be associated with them.

We have predicted 3 456 718 *Rho*-independent terminators, on average ~2.6 terminators per operon, suggesting alternative terminators for each operon, for all the 2072 organisms using the *TransTermHP* program (23), the best terminator predictor in the public domain, with the default parameters. All the terminators for each operon can be displayed both in an HTML table on the operon page and through the genome browser (see Figure 1C).

INTEGRATION OF CONSERVED OPERONS ACROSS BACTERIA

We have included the orthologous relationships among multi-gene operons across different bacterial genomes. Such information can be used for studies of operon evolution, such as elucidation of the life cycle of an operon (24). For two operons *a* and *b* in genomes *A* and *B*, respectively, we define a 'similarity score' between them as follows:

$$S(a,b) = \frac{|orth(a,b)|}{(|G(a)|+|G(b)|)/2}$$

where $G(a)$ and $G(b)$ denote the component genes in *a* and *b*, respectively; $orth(a,b)$ represents the orthologous gene pairs between *a* and *b* identified by our prediction program GOST (25) (see Figure 1D); and $|X|$ denotes the number of elements in *X*. Intuitively, the score = 1 if and only if all genes in *a* and *b* are one-to-one mapped to orthologous gene pairs; and the score = 0 if and only if no orthologous genes between *a* and *b* are detected. Generally, the higher the score, the higher percentage of genes in *a* and *b* are orthologous pairs. We consider a pair of operons *a* and *b* as *conserved* if $S(a,b)$ is at least 0.7. Using this cut-off, 6 975 454 conserved operon pairs have been identified among the 2072 genomes. For any specific operon, a user can retrieve its conserved operons across all the other 2071 genomes in DOOR 2.0 by selecting the relevant menu item on the browser.

A NEW WEB INTERFACE

The web interface of DOOR 2.0 is completely redesigned compared with the previous version. The new features of the interface include (i) an intuitive genome browser based on JBrowse Genome Browser (http://jbrowse.org) (26) that supports visualization of all the aforementioned data types related with operons along with a scrollable and zoomable chromosome for each organism; (ii) a new keyword-based Google-like search engine implemented using the Sphinx Open Source Search Server (http://sphinxsearch.com), through which a user can enter one or a few keywords to search for operons that have the specified attributes, e.g. *coli*, *lactose*, *NC_00913*, and can also formulate the search key as a complex query with Boolean operators (see online document on the DOOR 2.0 web server for examples); and (iii) an intuitive Web 2.0 HTML table (DataTables, https://datatables.net) that supports on-the-fly filtering, multi-column sorting, variable length pagination and asynchronous loading for large datasets.

ONLINE OPERON PREDICTION

DOOR 2.0 offers an intuitive high-performance web service for online operon prediction. A user can have operons predicted in a newly sequenced genome or any provided prokaryotic genome sequence by uploading three types of data into the server, including chromosomal DNA sequence (in *fna* format as used by the NCBI Genome FTP Server), protein sequence (*faa* format as

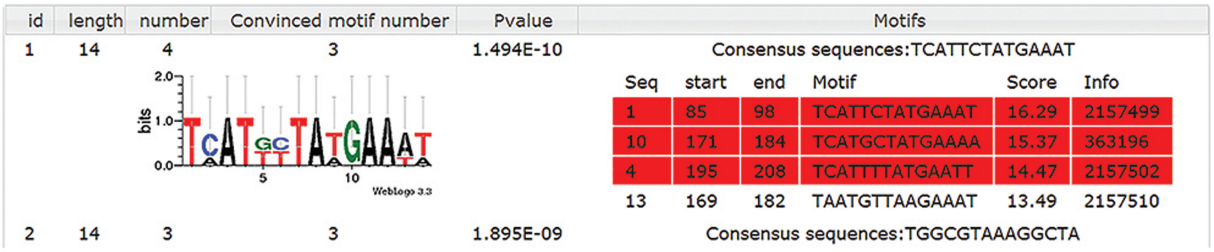

Figure 2. A screenshot of motif search results for a user-selected operon.

used by the NCBI Genome FTP Server) and gene location (*ptt* format as used by the NCBI Genome FTP Server). All the submitted jobs are put automatically into a job queue, which are executed in a 'first-in, first-served' manner on our computing cluster. Once the job is done, the user will be notified via email with links to the web pages containing the computational results. All the predicted operons are displayed in an intuitive HTML table and stored on the DOOR 2.0 server for half a year.

IMPLEMENTATION

DOOR 2.0 is implemented as a web portal server with a multi-layer architecture. The representation and the logic layers are implemented using the Web 2.0 technology (HTML5, CSS3 and Javascript language along with jQuery library) and PHP server-side scripting language. All data are stored in an optimized MySQL relational database. The keyword-based search engine is implemented based on the Sphinx Open Source Search Server (http://sphinxsearch.com), and the genome browser is implemented based on JBrowse Genome Browser (http://jbrowse.org) (26) and integrated into DOOR 2.0 using the *iframe* (inline frame) HTML tag. The web server runs on a Red Hat Enterprise Linux 6 box (16 Intel Xeon CPUs with 2.4 GHz and 16 GB memory), and automated operon prediction pipeline runs on the computing cluster server with >150 computing nodes (2 Intel Xeon CPUs with 3.06 GHz and 2.5 GB memory per node).

CONCLUDING REMARKS

Here we presented a new version of the DOOR operon database, DOOR 2.0. Although the previous version has been widely used (with over ~120 citations since its publication in 2009), we feel that it is time to develop and deploy a new version of the database to include all the prokaryotic genomes sequenced in the past few years, the available TU information experimentally validated or computationally derivable from RNA-seq data, as well as regulatory signals for each operon, which can be predicted based on comparative genome analysis. To best facilitate data retrieval, analysis and integrated applications of these data, we have developed a highly intuitive genome browser to support the visualization of these data types. With the high quality of our predicted operons, along with their regulatory signals and evolutionary conservation information, we believe that the new version of DOOR will continue to serve as a main source of operon data for the microbial research community.

ACKNOWLEDGEMENTS

The authors thank all the members of the CSBL Lab at UGA, especially Dr Wen-chi Chou (now a postdoc in Harvard University) for help of TU prediction. X.M. coordinated the implementation of the project as well as the writing of the article, designed and implemented the web server, and drafted the manuscript; Q.M. assisted in the coordination of the article, carried out the prediction of operons, TUs and conserved operons and helped in revising the manuscript; C.Z. participated in the implementation of the web server and prediction of operons; X.C. participated in the implementation of the web server and prediction of TU; H.Z. participated in the implementation of the web server; W.L. participated in the implementation of the web server and helped revising the manuscript; and Y.X. conceived the study and polished the manuscript. All authors read and approved the final manuscript.

FUNDING

National Science Foundation [DEB-0830024]; DOE BioEnergy Science Center [contract no. DE-PS02-717 06ER64304] [DOE 4000063512], which is supported by the Office of Biological and Environmental Research in the Department of Energy Office of Science. Funding for open access charge: National Science Foundation [DEB-0830024] and the DOE BioEnergy Science Center [contract no. DE-PS02-717 06ER64304] [DOE 4000063512].

Conflict of interest statement. None declared.

REFERENCES

1. Jacob,F., Perrin,D., Sanchez,C. and Monod,J. (1960) Operon: a group of genes with the expression coordinated by an operator. *C. R. Hebd. Seances Acad. Sci.*, **250**, 1727–1729.
2. Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) A probabilistic learning approach to whole-genome operon prediction. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 116–127.
3. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
4. Cho,B.K., Zengler,K., Qiu,Y., Park,Y.S., Knight,E.M., Barrett,C.L., Gao,Y. and Palsson,B.O. (2009) The transcription unit architecture of the *Escherichia coli* genome. *Nat. Biotechnol.*, **27**, 1043–1049.
5. Salgado,H., Peralta-Gil,M., Gama-Castro,S., Santos-Zavaleta,A., Muniz-Rascado,L., Garcia-Sotelo,J.S., Weiss,V., Solano-Lira,H., Martinez-Flores,I., Medina-Rivera,A. *et al.* (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.*, **41**, D203–D213.
6. Okuda,S., Katayama,T., Kawashima,S., Goto,S. and Kanehisa,M. (2006) ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res.*, **34**, D358–D362.
7. Sierro,N., Makita,Y., de Hoon,M. and Nakai,K. (2008) DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res.*, **36**, D93–D96.
8. Pertea,M., Ayanbule,K., Smedinghoff,M. and Salzberg,S.L. (2009) OperonDB: a comprehensive database of predicted operons in microbial genomes. *Nucleic Acids Res.*, **37**, D479–D482.
9. Taboada,B., Ciria,R., Martinez-Guerrero,C.E. and Merino,E. (2012) ProOpDB. *Nucleic Acids Res.*, **40**, D627–D631.
10. Mao,F., Dam,P., Chou,J., Olman,V. and Xu,Y. (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**, D459–D463.
11. Okuda,S. and Yoshizawa,A.C. (2011) ODB: a database for operon organizations, 2011 update. *Nucleic Acids Res.*, **39**, D552–D555.
12. Dam,P., Olman,V., Harris,K., Su,Z. and Xu,Y. (2007) Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res.*, **35**, 288–298.

13. Adhya,S. (2003) Suboperonic regulatory signals. *Sci. STKE*, **2003**, pe22.
14. Kodama,Y., Shumway,M. and Leinonen,R. (2012). International Nucleotide Sequence Database Collaboration. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
15. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
16. Kazakov,A.E., Cipriano,M.J., Novichkov,P.S., Minovitsky,S., Vinogradov,D.V., Arkin,A., Mironov,A.A., Gelfand,M.S. and Dubchak,I. (2007) RegTransBase–a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, D407–D412.
17. Ma,Q., Liu,B., Zhou,C., Yin,Y., Li,G. and Xu,Y. (2013) An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale. *Bioinformatics*, **29**, 2261–2268.
18. Li,G., Liu,B., Ma,Q. and Xu,Y. (2011) A new framework for identifying cis-regulatory motifs in prokaryotes. *Nucleic Acids Res.*, **39**, e42.
19. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
20. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
21. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
22. Watson,J.D., Baker,T.A., Bell,S.P., Gann,A., Levine,M. and Losick,R. (2013) *Molecular Biology of the Gene*. Pearson/Benjamin-Cummings Publishing Company, San Francisco, CA, USA.
23. Kingsford,C.L., Ayanbule,K. and Salzberg,S.L. (2007) Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol.*, **8**, R22.
24. Price,M.N., Arkin,A.P. and Alm,E.J. (2006) The life-cycle of operons. *PLoS Genet.*, **2**, e96.
25. Li,G., Ma,Q., Mao,X., Yin,Y., Zhu,X. and Xu,Y. (2011) Integration of sequence-similarity and functional association information can overcome intrinsic problems in orthology mapping across bacterial genomes. *Nucleic Acids Res.*, **39**, e150.
26. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.